Prendre Le Monde en main : choix d'architecture

Serge Heiden, Pierre Lafon, Gabriel Illouz, Benoît Habert, Serge Fleury, Helka Folch, Sophie Prévost

Prendre *Le Monde* en main : choix d'architecture

Proposition de communication pour RIAO 2000

B. Habert, G. Illouz, H. Folch, S. Fleury, S. Heiden, P. Lafon, S. Prévost
LIMSI & UMR 8503

16 novembre 1999

Résumé Le recours croissant aux « très grands corpus » pour améliorer les systèmes de Traitement Automatique des Langues (TAL) suppose de maîtriser l'homogénéité lexicale, morpho-syntaxique et syntaxique des données utilisées. Cela implique en amont le développement d'outils de calibrage de textes. Nous mettons en place de tels outils et la méthodologie associée dans le cadre de l'appel d'offres ELRA *Contribution à la réalisation de corpus du français contemporain*. Nous montrons sur les rubriques principales du journal *Le Monde* les premiers résultats de cette approche. Nous précisons les contraintes qui en résultent pour les chaînes de traitement de corpus, au regard des propositions existant dans le domaine.

Mots-clés acquisition de connaissances linguistiques, linguistiques de corpus, très grands corpus, typologie de textes.

1 Enjeux du « profilage » de corpus

1.1 Changement de cap en TAL

C'est désormais du recours aux « très grands corpus » qu'est attendue une amélioration significative des systèmes de TAL (Amstrong, 1994), via l'acquisition de connaissances linguistiques à la fois très vastes (en nombre d'entrées lexicales et de règles) et très détaillées (concernant les conditions syntaxiques d'emploi des mots ou leurs associations privilégiées, par exemple).

Alors que les données textuelles disponibles pour l'acquisition de connaissances lexicales, syntaxiques et sémantiques en TAL ont atteint des proportions volumineuses (comme les 100 millions de mots étiquetés du `BNC` – British National Corpus [1]), elles rassemblent parfois des composants extrêmement hétérogènes. Il en va ainsi des données de presse, comme les CD-ROM du *Monde*, qui sont souvent mises à contribution vu leur accessibilité.

1.2 Maîtriser les données utilisées en acquisition

Plusieurs études convergent pour rendre plausible l'hypothèse selon laquelle la fiabilité des traitements automatiques, dans différents domaines, dépendrait de l'homogénéité des

[1] Ce corpus – `http ://info.ox.ac.uk/bnc/` – mêle oral (10 %) et écrit (textes de fiction à partir de 1960 et textes « informatifs » à partir de 1975). Il répond à l'objectif de constituer un ensemble de données textuelles aux conditions de production et de réception définies avec précision et qui soient représentatives d'une grande variété de situations de communication. Cf. (Habert *et al.*, 1997, p. 147).

données en cause.

Étiquetage D. Biber a ainsi montré (Biber, 1993, p. 223) sur des corpus subdivisés en domaines (ici dans le LOB – *Lancaster-Oslo-Bergen*) [2] que la probabilité d'apparition d'une catégorie morpho-syntaxique donnée est fonction du domaine. D. Biber montre également (*ibid.*, p. 225) les différences dans l'enchaînement des probabilités des catégories morpho-syntaxiques d'un domaine à l'autre, ainsi que les différences dans les collocations (par exemple, pour *sure* et *certain*). Dans la perspective de la mise au point d'un étiqueteur probabiliste, se limiter à un seul des domaines biaiserait singulièrement les apprentissages. Les « additionner » sans plus de précaution aboutirait à des « moyennes » peu utilisables.

La précision des étiqueteurs participant à l'action d'évaluation GRACE (Adda *et al.*, 1999), mesurée par rapport au corpus de référence corrigé manuellement, manifeste également des variations significatives en fonction de la partie de ce corpus concernée (Illouz, 1999). Ce corpus de 100 000 mots rassemble en effet des extraits du *Monde* (2) et de textes littéraires : mémoires (2), romans (6), essais (2). Ainsi, un extrait de mémoires entraîne de fortes variations, positives et négatives, entre les étiqueteurs.

Parsage Sekine (Sekine, 1998) utilise 8 domaines du corpus BROWN (reportages, éditoriaux, « hobbies », « learned », fiction, western, romans sentimentaux). Il examine les performances, mesurées en rappel précision, d'un analyseur syntaxique probabiliste selon que l'apprentissage de la grammaire s'effectue sur le même domaine que celui du test, sur tous les domaines confondus, sur la partie *fiction* (fiction, western, romans sentimentaux) ou sur la partie *non-fiction* (reportages, éditoriaux, « hobbies », « learned ») (il appelle ces deux derniers regroupements des « classes »). Les performances vont en général dans l'ordre décroissant suivant : identité domaine d'apprentissage/de test, appartenance des domaines d'apprentissage/de test à la même « classe », apprentissage /test sur un corpus relevant de tous les domaines à la fois. Entraîner l'analyseur sur une classe (*fiction* par exemple) et l'utiliser sur l'autre classe (*non-fiction*) donne les résultats les plus mauvais.

Recherche d'information Karlgren (Karlgren, 1999, p. 159–161) utilise la portion du *Wall Street Journal* provenant du corpus TIPSTER [3] et les requêtes d'interrogation 202 à 300 de la campagne d'évaluation TREC (*Text Retrieval Conference*) assorties des jugements de pertinence sur les 74 516 articles en question [4], c'est-à-dire de l'indication que l'article est ou non une réponse correcte à une requête donnée. Elle mesure un certain nombre de caractéristiques stylistiques de chaque article : longueur moyenne des mots, proportion de mots longs, fréquence moyenne des mots, fréquence moyenne de mots capitalisés, proportion de nombres, pronoms personnels... À cette aune, il s'avère que les textes jugés pertinents diffèrent significativement des textes jugés non pertinents, et surtout que les textes les plus fréquemment retenus par les systèmes en compétition à TREC (qu'ils

[2] Ce corpus étiqueté a été conçu comme l'équivalent anglais de Brown, corpus étiqueté d'un million de mots au point en 1979 par W. Francis et H. Kučera, à l'université Brown (USA). Brown comprend 500 extraits de 2 000 occurrences chacun provenant de textes américains publiés en 1961 et relevant de 15 « genres » : reportage, écrits scientifiques et techniques, etc. Il a été soigneusement étiqueté. Par sa mise dans le domaine public, il a joué un rôle moteur dans le renouveau des études sur corpus. LOB comprend également 1 million de mots sélectionnés selon les mêmes critères mais à partir de textes anglais publiés en 1961.

[3] http ://www.tipster.org

[4] Ces articles proviennent des années 1990 à 1992. 2 039 sont pertinents pour au moins une requête. 35 289 ne sont pertinents pour aucune requête. Il reste 37 188 articles non jugés.

soient pertinents ou non) s'écartent également significativement des textes pour lesquels il n'y a pas de jugement de pertinence.

Comme le montrent les expériences faites en étiquetage, en parsage et en recherche d'information, en tant qu'échantillon de données langagières, un corpus est susceptible d'être à l'origine de deux types d'erreurs statistiques qui menacent les généralisations à partir de lui (Biber, 1993, p. 219–220) : « l'incertitude » (*random error*) et la « déformation » (*bias error*). L'incertitude survient quand un échantillon est trop petit pour représenter la population globale. Une déformation se produit quand une ou plusieurs caractéristiques d'un échantillon sont systématiquement différentes de celles de la population que cet échantillon a pour objectif de refléter [5]. C'est l'éventualité de ce type d'erreur qui enjoint de mieux connaître les paramètres proprement linguistiques du corpus.

2 Méthodologie et architecture de profilage

2.1 Définition

Nous appelons *profilage de textes* l'utilisation d'outils de calibrage donnant des indications sur l'emploi du vocabulaire, mais aussi de catégories et de patrons morpho-syntaxiques, etc., dans les parties d'un corpus, pour regrouper ces parties ensuite en sous-ensembles homogènes sur ces points.

L'optique, inductive, dans laquelle nous nous inscrivons consiste en effet à faire émerger *a posteriori* les types de textes – considérés comme des agglomérats fonctionnellement cohérents de traits linguistiques – grâce à un traitement statistique multidimensionnel de textes annotés. Cette optique constitue la ligne directrice des travaux de D. Biber (Biber, 1988)(Biber, 1995). Biber examine 67 traits linguistiques dans les 1 000 premiers mots de 481 textes d'anglais contemporain écrit et oral. Les traits étudiés ressortissent à 16 catégories distinctes (marqueurs de temps et d'aspect, questions, passifs, modaux...). Ils sont identifiés automatiquement sur la base d'un premier étiquetage morpho-syntaxique. La statistique multidimensionnelle permet d'obtenir des pôles multiples, positifs et négatifs, correspondant à des constellations de traits linguistiques corrélés. Ces pôles constituent deux à deux des dimensions textuelles. Chaque texte, par son emploi des traits linguistiques retenus, se situe en un point déterminé de l'espace à n dimensions issu de l'analyse. Les techniques de classification automatique permettent alors de regrouper les textes en fonction de leurs coordonnées sur ces dimensions. Les types de textes qui en résultent ne recoupent directement ni les « genres » des données de départ ni les registres intuitivement distingués.

Des études contrastives des performances, sur ces types de textes, de différents outils de TAL permettront de tester leur robustesse face à cette variation langagière, de déterminer les types auxquels chacun d'entre eux est spécifiquement adapté et d'ajuster les traitements. À l'inverse, les outils de calibrage pourront positionner un nouveau texte par rapport aux regroupements déjà obtenus et ainsi aideront à choisir les traitements les plus adaptés.

Nous mettons en place, dans le cadre du projet `TyPTex` (*Typage et Profilage de Textes*)

[5] Y. Wilks et R. Gaizauskas (Wilks & Gaizauskas, 1999, p. 198) commentent ainsi le privilège *de facto* donné au *Wall Street Journal* dans les travaux récents en recherche d'information : « ... le sur-entraînement sur un seul type de texte peut avoir eu des effets profonds mais pas encore mesurés sur le champ : le traitement linguistique du *Wall Street Journal* s'est certainement amélioré, mais il n'est pas sûr qu'il puisse servir de pierre de Rosette pour défaire les secrets de tous les textes du monde ! »

commun au LIMSI et à l'UMR 8503 et soutenu financièrement par l'ELRA (*European Language Resources Association*) dans le cadre de l'appel d'offres *Contribution à la réalisation de corpus du français contemporain*, une méthodologie permettant de tester et d'étendre les propositions de Biber, en utilisant en particulier les acquis pour le français de (Sueur, 1982) et de (Bronckart *et al.*, 1985). Les trois projets retenus dans cet appel d'offres utiliseront un corpus commun de 5 millions de mots, dont un million provenant du journal *Le Monde* et constituant un sous-ensemble de la partie *Presse* du corpus PAROLE (cf. infra, section 3.1).

2.2 Organisation globale pour le projet TYPTEX

Comme le montre la figure 1, p. 4, on dispose au départ d'une base de textes. Chacun comprend un cartouche de description suivant les recommandations de la TEI (*Text Encoding Initiative*) (Dunlop, 1995). Les critères d'une requête d'extraction ou d'une sélection aboutissent à un corpus, c'est-à-dire un ensemble de textes rassemblés en fonction d'une recherche ou d'une application déterminée. Chacun de ces textes est soumis à un *étiquetage morpho-syntaxique*, qui permet d'associer à chaque mot ou unité polylexicale un lemme, une partie du discours et des indications morphosyntaxiques plus fines. Le *marquage typologique* utilise alors l'ensemble de ces informations et les **remplace** par de nouvelles catégories, correspondant aux traits dont on veut étudier la distribution. Le corpus marqué (et éventuellement corrigé par le biais de `CorTecs` (Heiden *et al.*, 1998)) est alors soumis à des logiciels d'analyse textuelle. En particulier, on construit la matrice des fréquences de chaque trait dans chaque texte. Cette matrice sert tant à la recherche optimale de traits pertinents à une opposition, qu'à la classification inductive ou supervisée.

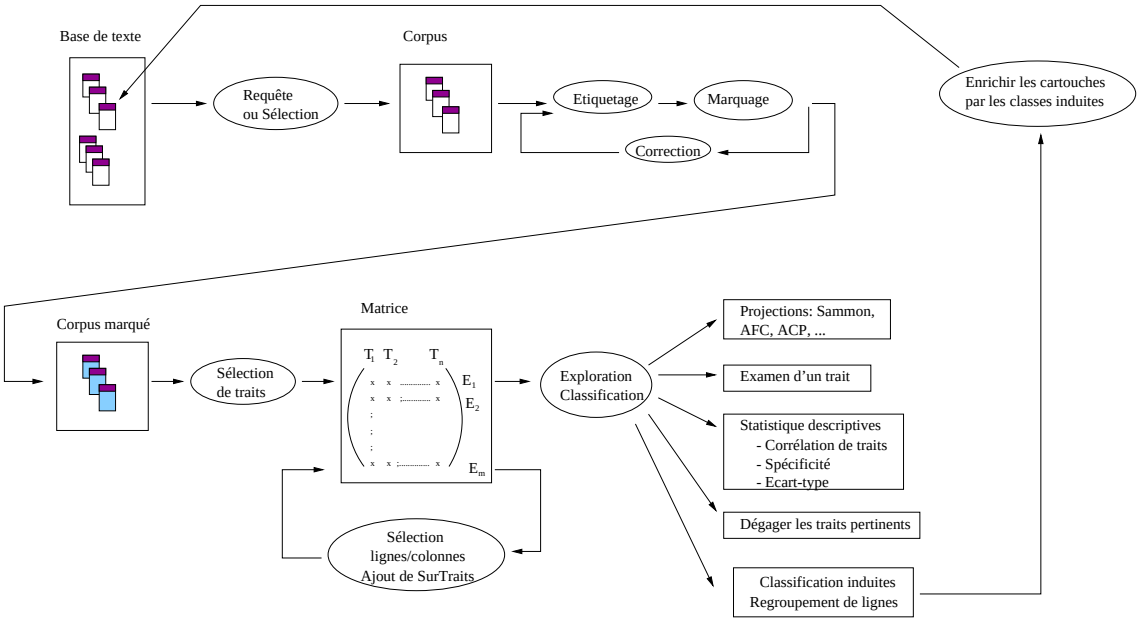

FIG. 1: Architecture de profilage de textes.

Nous avons utilisé pour l'étiquetage `Sylex-Base` (Ingenia, 1995), étiqueteur/analyseur basé sur le travail de P. Constant (Constant, 1991). Robuste, il a obtenu des résultats satisfaisants lors de l'action d'évaluation d'étiqueteurs `GRACE`. Le marquage opéré en aval, dans

l'immédiat limité, porte par exemple sur les embrayeurs, les modalités, les présentatifs, l'usage des temps, le passif, certaines classes d'adverbes (négation, degré...) et de déterminants, etc. On garde la catégorie (partie du discours) des mots ou unités polylexicales qui n'ont pas été autrement marqués.

3 Expérience : prendre *Le Monde* en main...

3.1 Données utilisées et expériences précédentes

Le corpus réalisé par G. Vignaux (INaLF) et B. Habert dans le cadre du projet européen PAROLE comprend une partie *Presse* de 14 millions de mots provenant par choix aléatoire de numéros entiers du journal *Le Monde* parmi ceux des années 1987, 1989, 1991, 1993 et 1995 (Naulleau, 1998).

L'étude (Illouz *et al.*, 1999) menée sur les 6 rubriques les plus importantes en volume – ART [6] (arts, médias, spectacles), ECO(nomie), EMS (? éducation, médecine, société), ETR(ranger), ING (? information générale : sport, faits divers), POL(itique) – de cette partie *Presse* montrait des écarts significatifs entre ces rubriques à la fois pour le vocabulaire utilisé et pour les catégories syntaxiques qui y sont privilégiées [7].

La série d'expériences présente les traitements que permet l'architecture TypTex. Elle porte sur les mêmes sections. Mais elle s'en tient aux articles comprenant de 1 000 à 2 000 mots : on évite de comparer des articles de taille trop dissemblable [8]. Ce sous-corpus totalise 2 160 071 occurrences. Chacun des articles a été étiqueté par `Sylex-Base`. Les étiquettes obtenues ont été remplacées par des catégories typologiques (cf. supra, p. 2.2).

3.2 Des rubriques distinctes *mais* des articles proches

Nous avons utilisé la méthode des spécificités (Lafon, 1980) pour dégager les sur et sous-emplois significatifs d'une catégorie dans une rubrique par rapport à sa répartition dans les 6 rubriques. Dans le tableau 1, le + indique un sur-emploi, le -, un sous-emploi, l'indication numérique qui suit donne l'ordre de grandeur de la probabilité de ce sur- ou sous-emploi. Des quelques deux cents traits actuellement marqués, nous en avons retenu une quarantaine, en deux sous-ensembles. Le premier comprend les éléments-outils assurant l'organisation du discours et de la phrase, le second rassemble les catégories ouvertes : noms, adjectifs, temps verbaux...

L'outillage grammatical dans ART manifeste une structuration textuelle expressive voire affective : sur-emplois du *Point d'exclamation*, du *Point d'interrogation*, des présentatifs ($C'est_{Indicatif\,présent}$, $Il\ y\ a_{Indicatif\,présent}$), des *Adverbes de degré*, ainsi que des pronoms, en parculier *Pro. pers. 1^e pers. sing.* et *Pro. pers. 2^e pers. plur.* ECO est presque à l'opposé. POL fait appel fortement aux *Deux points*, au *Subordonnant « que »*, ainsi qu'aux *Adverbes de négation* : marques de discours rapporté (direct ou indirect) et de polémiques ?

[6] Ce sont les classifications utilisées par la rédaction du journal *Le Monde* qui sont reprises dans les champs signalétiques du corpus PAROLE. On ne dispose pas toujours de la signification des libellés, d'où les points d'interrogation.

[7] Pour le vocabulaire, ont été étudiés les 15 438 articles totalisant 7 millions de mots et relevant de ces rubriques. Pour les catégories syntaxiques, ont été examinés 241 484 mots, provenant de 7 numéros de septembre 1987, qui ont été extraits de l'ensemble précédent, étiquetés automatiquement et corrigés manuellement pour la partie du discours, toujours dans le cadre de PAROLE.

[8] Les articles vont de 13 à 5 202 mots, avec une moyenne de 455.

Les *Adjectifs* opposent ART, ECO, EMS et *ETR* qui les sur-emploient à ING et POL qui les évitent. Le sur-emploi important des *Nom propre* dans POL est probablement favorisé par la présence de résultats électoraux. Les nominalisations (« la prise en main » *versus* « *X* prend en main ») – un même effacement des agents donc – rapprochent ECO et POL. Une certaine distance sur ce qui est rapporté – via le sur-emploi de *Pouvoir Conditionnel présent* – rapproche ECO et ETR, tandis que POL manifeste l'obligation [9] (*Il faut* et *Devoir* à l'*Indicatif présent*) et qu'ART fait coexister sur- et sous-emplois pour ces modalisations. Les temps sont sollicités diversement selon les sections : ART privilégie l'opposition *Présent* / *Passé simple* (et sous-emploie le *Passé composé*) ; ECO favorise participes, *Futur* et *Conditionnel présent* (s'agit-il de l'annonce « prudente » de certaines nouvelles ?) ; EMS fait fortement appel au *Futur* ; ETR sur-emploie les temps du passé comme ING qui écarte par ailleurs *Futur* et *Présent* ; POL marque un déficit net pour tous les temps sauf le *Présent*. Le passif oppose les sur-emplois d'ECO et EMS aux sous-emplois d'ART et POL, tandis qu'ETR et ING couplent sur- et sous-emplois.

À cet écart net entre les rubriques prises comme « en bloc », s'oppose l'homogénéité stylistique globale des articles pris un à un. En effet, lorsque chaque article est représenté par un vecteur constitué du nombre d'occurrences des deux cents traits retenus à des fins typologiques, les contrastes entre sections sont nettement moins marqué comme la figure 4. en témoignera. Nous avons donc conçu des outils pour observer plus finement les traits selon les classifications utilisées par la rédaction du journal *Le Monde*. Le premier étant la possibilité de relativiser l'ensemble des occurences de traits à un trait donné. Dans la suite, cette opération a été effectuée avec le nombre de mots.

3.3 Distributions des fréquences de traits

Prenons par exemple, les traits *Virgule* et *Nom Propre*, notés comme distinctifs par l'analyse des spécificités pour chaque section. Leurs distributions sont représentées en figures 2a et 2b. Celles-ci donnent le minimum, le premier quartile, la médiane, le troisième quartile, et le maximum, qui sont représentés par les cinq traits horizontaux.

On remarque que bien que les traits soient distribués différemment selon les sections, l'espace des valeurs pour les traits sont globalement les mêmes, du moins, il existe pour chaque classe de textes ayant des valeurs identiques ou proches pour un traits donné. Pour qu'un trait soit pertinent pour une classification à lui seul il faudrait qu'il partitionne l'ensemble des valeurs pour chaque classe. Nous l'avons donc vérifié pour chaque trait.

Après l'étude menée pour chacun des 200 traits, nous n'avons obtenu aucun partitionnement en classe. Nous avons alors exploré la corrélation des traits deux à deux, mais de façon non-exaustive (environ 20000 combinaisons possibles).

3.4 Exploration de corrélations de deux traits

Reprenons les deux traits de l'exemple précédent et examinons leur corrélation, ce qui correspond à étudier des nuages de points(articles), par classes (caractères a,b,c,d,e,f). Dans la figure 3, nous présentons les résultats pour toutes les sections. on note en haut au centre un regroupement de textes POL(caractère f), correspondant à des résultats d'élection. Le reste du graphique étant peu lisible, nous présentons aussi les résultats pour les seules classes ART et ECO (figure 3 b). Dans ce dernier graphique, on observe bien que les

[9] Le retour aux contextes serait nécessaire pour voir si c'est plutôt l'interprétation d'éventualité de *Devoir* qui apparaît dans ECO et celle de nécessité dans POL.

Trait	ART	ECO	EMS	ETR	ING	POL
Virgule	+**E51**	-*E51*	-*E31*	-*E51*		+**E51**
Deux points	+**E17**	-*E18*	-*E03*	-*E02*	+**E04**	+**E02**
Guillemets	-*E51*	-*E51*	-*E08*	+**E51**		+**E20**
Point d'exclamation	+**E10**	-*E09*	+**E02**	-*E07*		+**E07**
Point d'interrogation	+**E08**	-*E03*	+**E08**	-*E10*	+**E02**	
Point final	+**E17**		+**E03**	-*E27*	+**E11**	
Subordonnant « que »	-*E12*	-*E07*	-*E03*	+**E06**		+**E11**
Sub. relatifs	+**E12**	-*E06*	-*E02*	+**E08**	-*E02*	-*E08*
$C'est_{Indicatif\,présent}$	+**E32**	-*E13*		-*E06*	+**E02**	
$Il\ y\ a_{Indicatif\,présent}$	+**E08**	-*E02*			-*E02*	-*E03*
Coordonnants	+**E51**	-*E03*	+**E05**			-*E39*
Prépositions	-*E18*	+**E51**	+**E06**	+**E38**		-*E51*
Adverbes	+**E04**	+**E10**	+**E02**	+**E06**		-*E48*
Adv. de degré (« très »,...)	+**E09**	+**E13**			-*E03*	-*E28*
Adv. de négation	-*E02*	-*E21*	+**E02**	+**E03**	+**E08**	+**E03**
Pronom	+**E20**	-*E22*		-*E05*	+**E11**	+**E03**
Pro. pers. 1^e *pers. sing.*	+**E51**	-*E51*	+**E06**	-*E41*	+**E08**	-*E02*
Pro. pers. 2^e *pers. sing.*	+**E04**	-*E05*			+**E02**	
Pro. pers. 1^e *pers. plur.*	-*E04*	-*E03*	-*E02*	+**E16**		-*E03*
Pro. pers. 2^e *pers. plur.*	+**E18**	-*E11*	+**E02**	-*E09*	+**E02**	+**E03**
Pro. personnel 3^e *pers.*	+**E05**	-*E25*			+**E06**	+**E03**
Article défini	-*E47*	+**E37**		+**E51**	-*E04*	-*E51*
Article indéfini	+**E51**			+**E03**		-*E51*
Prép.+Art. déf. (« des »,...)			+**E04**	+**E04**		-*E20*
Adjectifs	+**E04**	+**E22**	+**E02**	+**E51**	-*E21*	-*E51*
Nombres cardinaux	-*E25*		+**E18**	-*E51*	+**E02**	+**E51**
Nom_{commun}	-*E02*	+**E37**	+**E16**	-*E04*	+**E06**	-*E44*
$Nom_{nominalisation}$	-*E51*	+**E51**			-*E17*	+**E02**
Nom_{propre}	+**E27**	-*E51*	-*E51*	+**E04**		+**E51**
$Devoir_{Indicatif présent}$	-*E02*	+**E04**		-*E05*		+**E04**
$Il\ faut_{Indicatif\,présent}$	+**E02**			-*E09*		+**E05**
$Pouvoir_{Conditionnel\,présent}$	-*E03*	+**E04**		+**E02**		-*E03*
Indicatif futur	-*E07*	+**E15**	+**E03**		-*E03*	-*E04*
Indicatif imparfait		-*E32*	-*E03*	+**E14**	+**E24**	-*E03*
Indicatif présent	+**E27**				-*E03*	-*E18*
Conditionnel	-*E06*	+**E02**		+**E02**		
Passé composé	-*E19*	-*E05*	-*E06*	+**E38**	+**E03**	-*E03*
Passé simple	+**E13**	-*E19*		+**E02**	+**E31**	-*E14*
Participe passé		+**E04**		+**E13**	+**E06**	-*E36*
Participe présent	-*E10*	+**E04**	-*E02*	+**E02**	+**E02**	
Futur Passif	-*E03*	+**E08**	+**E05**	-*E02*	-*E02*	-*E03*
Infinitif Passif	-*E08*	+**E04**	+**E04**			-*E02*
PasséComposé Passif	-*E08*			+**E17**	+**E03**	-*E11*
Présent Passif		+**E03**	+**E02**	+**E03**		-*E16*

TAB. 1: Oppositions stylistiques entre les 6 sections principales du *Monde*.

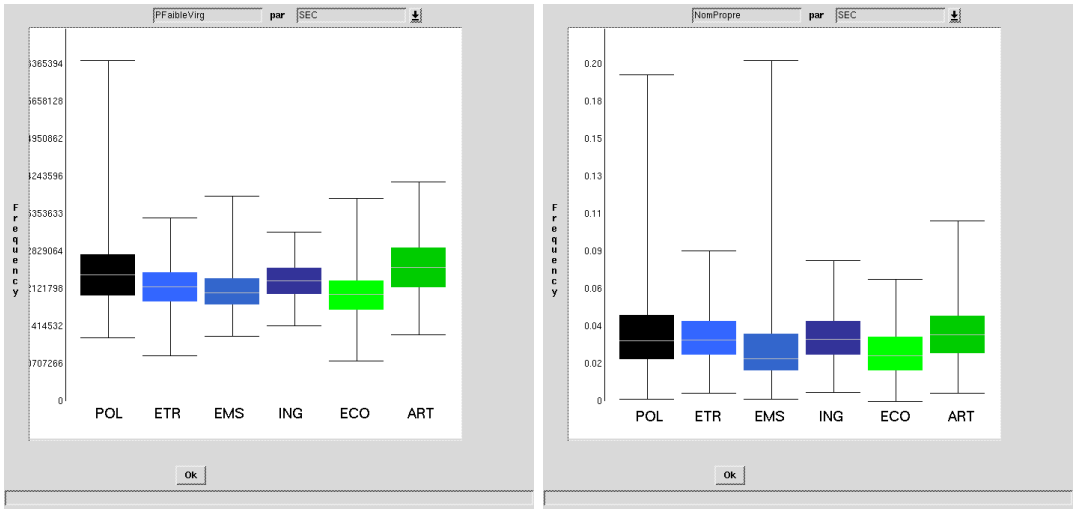

FIG. 2: Distibutions : a.Virgule, b.Nom Propre

centres de gravité des deux classes sont différents, confirmant l'analyse des spécificités, mais la frontière reste floue et ne permet pas d'induire de classification.

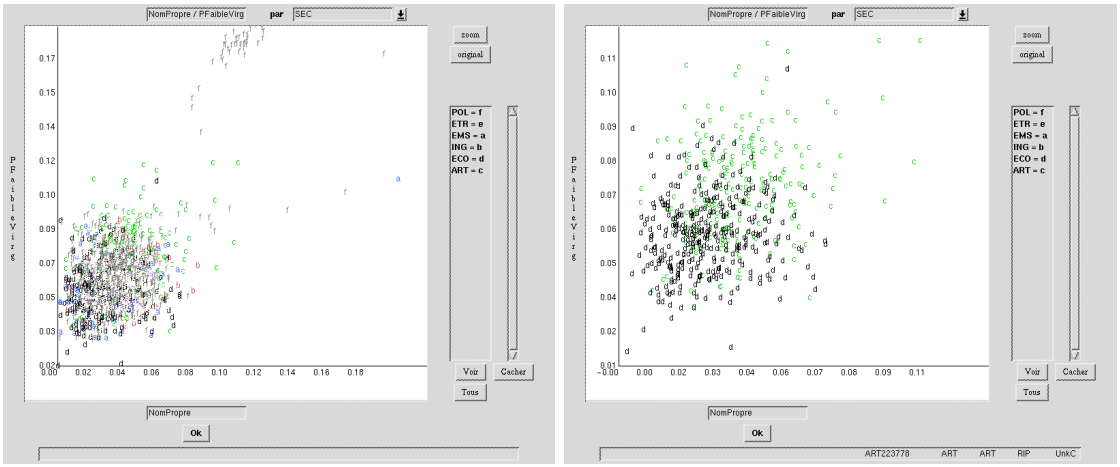

FIG. 3: Combinaison Virgule/Nom Propre, a.Toutes sections b.ART/ECO

Nombre de couples ont ainsi été étudié (par exemple nom/adjectif pour vérifier si l'adjectivisation est en proportion constante). Néanmoins, il n'a pu être mis en évidence des couples fortement discriminants. Nous avons alors utilisé des analyses multi-dimensionnelles, la projection de Sammon pour l'instant (l'AFC est en cours d'implémentation dans le module d'analyse de l'architecture TYPTEX).

3.5 analyses multi-dimensionnelles des distributions de traits

Les vecteurs de traits correspondant à chaque article sont explorés à l'aide de la méthode de Sammon (Sammon, 1969) qui, partant d'un nuage de dimension n, projette les données dans un espace de dimension k ($k << n$), avec la propriété de conserver au mieux les distances existantes dans l'espace de départ. Dans la figure 4, les points correspon-

dant aux articles ne se rassemblent pas dans des zones distinctes selon la rubrique dont ils proviennent. Dans (Illouz *et al.*, 1999), en utilisant les 108 formes supérieures à 500 occurrences comme traits, des échantillons de 10 000 mots par sections, et la même méthode de projection, on distinguait au contraire nettement le regroupement des articles selon les rubriques. Peut être que la fenêtre d'observation (de 1000 à 2000 mots) est ici trop fine pour obtenir des effets observables.

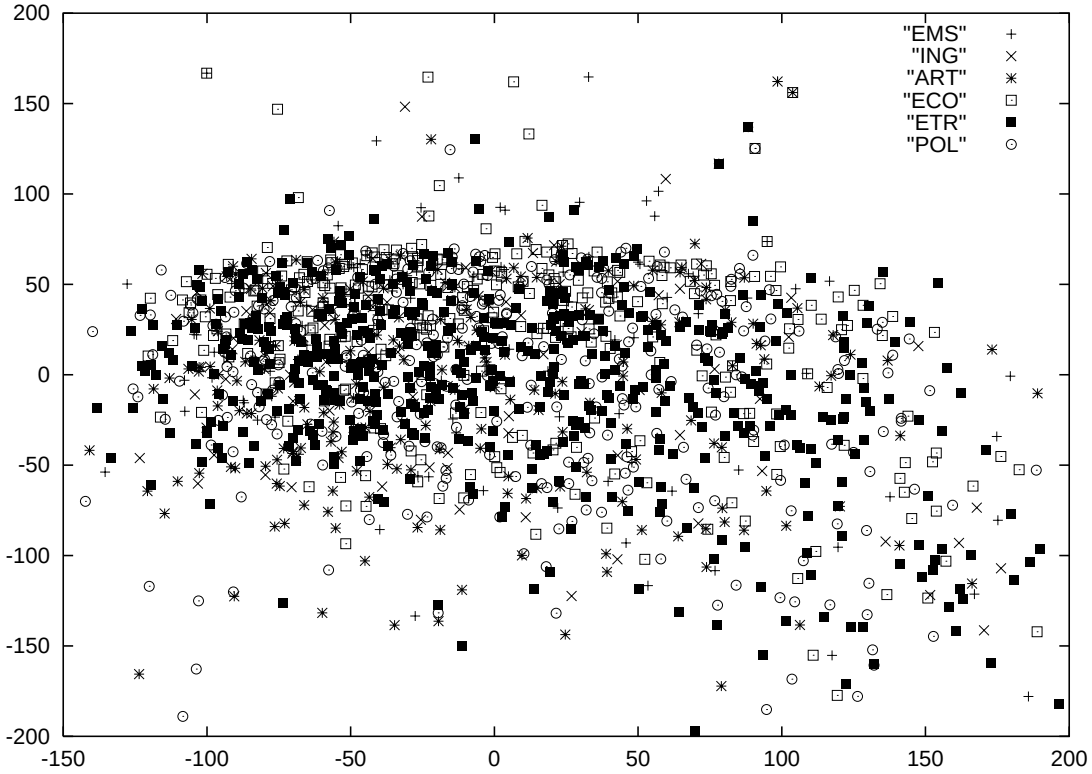

FIG. 4: Emploi des traits stylistiques par les articles des 6 sections.

Cette contradiction entre l'écart global entre les rubriques et les recoupements entre les articles pris individuellement ne donne par contre aucune indication sur l'existence dans le corpus étudié de types de textes définis comme corrélations entre traits. C'est le recours à la statistique multidimensionnelle qui permettra d'examiner cette hypothèse.

Cette expérience de marquage typologique permet néanmoins de revenir de manière critique sur les traits utilisés. Ils peuvent d'abord être trop « fins » et déboucher sur un éparpillement d'occurrences rendant impalpables les contrastes. C'est le cas dans la grille utilisée actuellement pour les temps des verbes : la catégorie verbale est « éclatée » en une cinquantaine de traits, dont la plupart totalisent un nombre limité d'occurrences. On ne dispose ainsi d'aucune prise sur le verbe dans son ensemble ni sur la manière dont cette catégorie est sollicitée selon les rubriques ou selon les articles. On ne sait par exemple pas si le sous-emploi des noms dans POL s'accompagne d'un sur-emploi du verbe, comme on l'observait pour un sous-ensemble du même corpus dans (Illouz *et al.*, 1999), ce qui serait cohérent avec le sous-emploi de tous les temps verbaux visibles dans le tableau 1. À l'inverse, certains traits sont trop grossiers et cachent probablement des oppositions effectives. Il en va ainsi de *nombres cardinaux* qui regroupe les indications de quantité, mais aussi les

dates, que l'on gagnerait probablement à distinguer. On souhaiterait en fait manipuler des traits structurées de manière à pouvoir utiliser tout ou partie des informations correspondantes [10]. Ainsi, disposer de l'étiquette {catégorie=nom, type=commun, genre=masculin, nombre=singulier...} permet de garder des sous-ensembles comme {catégorie=nom}, {catégorie=nom, type=commun}, voire {genre=masculin}. Utiliser des structures de traits du type de celles employées dans les grammaires d'unification permettrait de modéliser plus strictement les informations issues du marquage, dans l'esprit par exemple de (Gazdar *et al.*, 1990) ainsi que les opérations dont elles sont passibles [11].

Dans cet esprit, nous avons testé la possibilité de sommer des traits et de créer ce que nous avons appelé un *surtrait*. Ainsi à partir des traits élémentaires, nous avons construit les surtraits *Forme Verbale*, *Adverbe*, *Substantif*, *Pronom Personne*, et *Déterminant*. Puis, nous les avons analysé par une projection de Sammon sur les classes. Ces classes n'étant toujours pas significativement distinctives, nous avons étudié la classification par genre. Celle-ci bien que moins présente dans *Le Monde* [12] est néanmoins plus fine pour observer certaines distinctions de style est mettre au point notre profileur. La projection de Sammon correspondante est présentée en figure 5, seules les classes les plus pertinentes et en effectif suffisant ayant été gardées. On voit ainsi apparaître un regroupement des textes "marchés" qui occupe une place différente des textes "nécrologie" et "chronologie", qui eux ont des *comportements stylistiques* comparables.

On le voit, il faut pouvoir regrouper des traits pour un contraste, en éclater d'autres, voire recommencer sur certains points l'étiquetage et le marquage. La tâche de profilage de corpus impose donc une flexibilité dans le traitement de corpus qui introduit des contraintes sur les architectures logicielles à utiliser.

4 Évaluation des architectures de traitements de corpus

Paradoxalement, c'est l'échange des données langagières, brutes ou annotées, qui a concentré l'essentiel des initiatives de standardisation, dans le cadre de la TEI (Ide & Véronis, 1995). Plusieurs voies ont par contre été explorées pour les chaînes de traitement de corpus. Deux problèmes se posent en effet (McKelvie *et al.*, 1997) : comment traiter des corpus volumineux et assortis d'annotations complexes (étiquetage morpho-syntaxique ou sémantique, arbres syntaxiques, marques de co-référence...) et parfois contradictoires [13] ; comment articuler des composants logiciels de manière modulaire ? Nous présentons en section 4.1 trois architectures que nous avons testées et leurs réponses à ces problèmes et en section 4.2 nos propres choix en fonction de la tâche qui est la nôtre.

4.1 Trois architectures : GATE, IMS-CWB, LT NSL

Nous avons testé les trois architectures suivantes [14] :

[10] C'est l'approche de (Habert & Salem, 1995).

[11] Celles-ci ne sont pas assez contraintes actuellement. Ainsi, les mots marqués $Nom_{nominalisation}$ devrait être englobés dans la catégorie Nom_{commun}, ce qui n'est pas le cas actuellement.

[12] seul 15% des articles reçoivent un genre.

[13] Deux étiqueteurs morpho-syntaxiques appliqués à un même corpus peuvent opérer deux segmentations en unités lexicales différentes et diverger par ailleurs sur l'analyse faite de certains segments.

[14] Au LIMSI, G. Illouz a mis en œuvre GATE pour tester les conséquences du remplacement, au sein d'un module (comme l'étiquetage), d'un logiciel par un autre. À l'UMR 8503, S. Heiden a intégré le moteur de requêtes CQP d'IMS-CWB et l'architecture SGML de LT NSL comme outils de requêtes évoluées pour le logiciel d'exploration de données textuelles qu'il a développé : Lexploreur (Heiden, 1999).

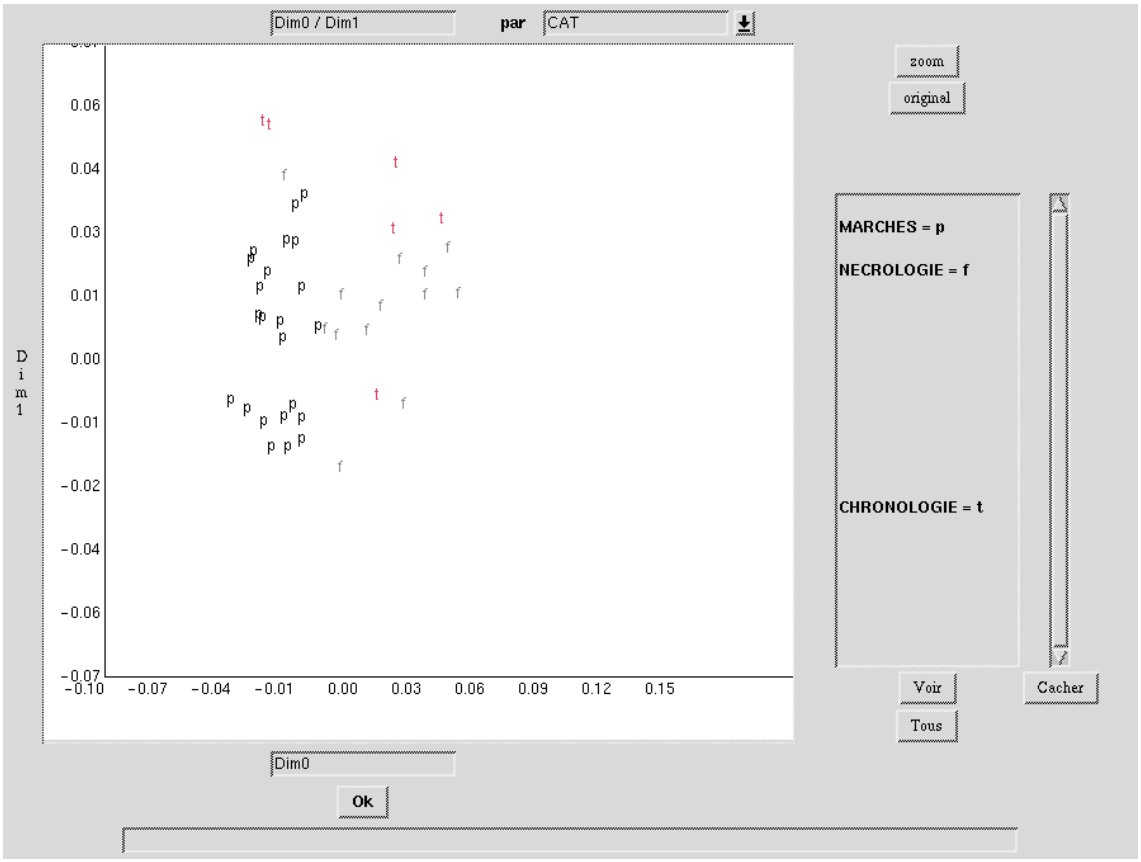

FIG. 5: Projection de Sammon sur les genres.

GATE (Wilks & Gaizauskas, 1999) Cette architecture est basée sur la nécessité de faire collaborer des modules de TALN hétérogènes pour le développement de systèmes complexes. Les annotations sont stockées séparément des données de départ auxquelles elles renvoient. L'interface graphique facilite la mise en relation des modules (qui nécessite cependant l'écriture d'outils de mise en correspondance des formats d'entrée et de sortie des modules avec le format pivot [15]) ainsi que l'exploration de diverses combinaisons de modules existants.

IMS Corpus Workbench (Christ, 1994) Cet atelier a été développé autour d'un moteur de recherche pour l'étude de corpus balisé. Les données textuelles, qui peuvent être assorties d'autant d'annotations que nécessaire, sont considérées comme une base de données. Cette base est stockée et indexée de sorte que les requêtes – en termes d'expressions régulières sur tout ou partie des annotations ou des suites d'annotations – puissent recevoir une réponse rapide. Cette architecture convient tout particulièrement à l'utilisation efficace d'un corpus dont l'annotation est « stabilisée ».

LT NSL (McKelvie *et al.*, 1997) LT NSL généralise l'approche par filtres successifs (*pipelines*) d'UNIX. Mais les données, à toutes les étapes de leur traitement, sont balisées en SGML. L'arbre ou la succession d'événements que constitue un document SGML une fois parsé fournissent un contexte aussi précis que souhaité pour les re-

[15] Chaque annotation atomique est constituée de son type – mot, syntagme, phrase...–, de ses attributs, et de ses positions de départ et d'arrivée dans les données originelles.

quêtes. Cette architecture permet d'expérimenter différents types d'annotations tout en garantissant la correction formelle des différents états des données et un parsage optimisé du flux SGML.

Deux solutions s'offrent donc pour l'utilisation simultanée d'annotations multiples : stocker les annotations en un document unique (IMS-CWB) *versus* répartir les annotations (GATE). La première facilite les accès ultérieurs au document et les mises en relation directes des différents niveaux d'annotation. La seconde s'impose quand les annotations divergent. Elle facilite l'articulation d'un grand nombre d'annotations simultanées. Par ailleurs, l'enchaînement de composants logiciels peut s'effectuer par l'utilisation d'un format pivot pour le lien entre deux modules (GATE) – chaque module restant par ailleurs « maître chez lui » – ou par le recours de chacun des modules à un format unique (LT NSL). La première solution favorise l'utilisation conjointe de modules hétérogènes, la seconde l'homogénéité des traitements.

4.2 Contraintes liées au profilage

Nous détaillons, pour chacune des étapes du profilage, les contraintes à prendre en compte et leur compatibilité avec les trois architectures qui viennent d'être présentées.

Varier les corpus À partir d'un ensemble de textes, de nombreux corpus distincts peuvent être constitués, en fonction des critères d'extraction choisis. H. Folch a developpé [16] pour ce faire un outil de création de corpus à partir d'une requête utilisant les champs signalétiques présents dans les cartouches de la base textuelle. Aboutir à des types de textes stables suppose de constituer de multiples corpus pour examiner la stabilité des regroupements opérés par les outils de profilage [17]. GATE et IMS-CWB, qui « figent » la collection de textes utilisés, sont mal adaptés à cette nécessité.

Des traits évolutifs Les traits utilisés pour profiler les textes sont mouvants. En premier lieu, leur liste n'est pas close. Ils ressortissent à des niveaux différents :

caractères employés ponctuation, majuscules et chiffres en particulier (Illouz, 1999) ;

ensembles lexicaux clos catégories de « mots-outils » (Brunet, 1981) (Biber, 1988)(Illouz *et al.*, 1999) ;

classes lexicales ouvertes nom propre / nom commun...(Biber, 1988) (Illouz *et al.*, 1999) ;

catégories typologiques plus fines (Sueur, 1982)(Bronckart *et al.*, 1985) (Biber, 1988) ;

organisation du texte titrage, présence d'images, de tableaux (Karlgren, 1999).

En second lieu, les outils d'étiquetage morpho-syntaxiques utilisés en amont peuvent rendre plus ou moins aisé le repérage de tel ou tel d'entre eux, voire pousser à abandonner certains (passifs sans agents, par exemple). On peut par ailleurs recourir à plusieurs d'entre eux.

[16] Dans le cadre du projet SCRIPTORIUM de veille sociale interne à la Direction des Études et Recherches d'EDF. Ce projet, qui fait l'objet d'un contrat pour la période 1997-2000 entre la Direction des Etudes et Recherches d'EDF et l'ENS de Fontenay/Saint-Cloud, comprend la constitution d'un corpus de 20 millions de mots. Les documents rassemblés sont extrêmement variés tant pour le format que pour le type de données langagières : tracts, extraits de livre, presse syndicale, presse d'entreprise, comptes-rendus de comités d'entreprise, transcriptions de messages syndicaux enregistrés, etc.

[17] Comme l'a fait Biber dans (Biber, 1995) où il compare les résultats obtenus pour quatre langues distinctes.

Enfin, il reste à déterminer empiriquement les traits qui sont effectivement discriminants et qui permettent d'obtenir des types de textes nettement définis. Cette instabilité exclut dans la phase de mise au point le recours à IMS-CWB.

Des traitements statistiques multiples Ils s'organisent en deux volets. Le premier a pour objectif l'exploration des corrélations significatives de traits linguistiques (ACP, AFC, Sammon). Il suppose de pouvoir observer un trait ou un petit groupe de traits pour déterminer leur pertinence vis-à-vis d'une classification. Il doit permettre de manipuler des traits qui ne relèvent pas forcément des mêmes lois de probabilité (Karlgren, 1999, p. 153), ce qui implique de pouvoir visualiser les textes comme points dans un espace en pouvant changer de point de vue, de classification. Le deuxième volet relève de l'apprentissage supervisé. Il revient à pouvoir situer un texte donné dans une classification pré-existante (via C4.5 de Quinlan, par exemple).

Les divers outils évoqués sont dispersés dans différentes communautés (analyse de données et apprentissage automatique) et, par conséquent, difficiles à utiliser simultanément. GATE permet en principe leur articulation, mais conduit à « envelopper » les données produites pour assurer l'interopérabilité des traitements utilisés. Pour simplifier ce problème, nous avons choisi de définir une matrice de contingence unique où les individus sont les textes et les variables les traits ce qui permet de fournir à chaque outil les données nécessaires [18].

Un lien permanent textes de départ / profilages Les corrélations de traits issus de traitements multidimensionnels résistent souvent à l'interprétation (Karlgren, 1999, p. 157), comme l'a montré l'examen du tableau 1. Pouvoir examiner le fonctionnement de ces traits dans le contexte des textes réunis lors de la constitution du corpus est crucial pour contrôler les interprétations proposées. Par ailleurs, certains des résultats obtenus par les outils de profilage sont autant de renseignements signalétiques à ajouter aux cartouches des textes de départ. Aucune des trois architectures évoquées *supra* ne permet pourtant de bien prendre en compte ce lien.

5 D'un prototype de profileur à une architecture générique

Nous disposons d'un prototype de « profileur » de corpus. Il permet globalement les traitements de la figure 1, p. 4 :

Constitution de corpus Un moteur de requêtes utilise les champs signalétiques de la base textuelle pour constituer des corpus raisonnés ;

Marquage typologique Transformation des résultats d'un étiqueteur morpho-syntaxique pour mettre l'accent sur des fonctionnements langagiers constitutifs de types de textes ;

Examen des traits deux à deux Visualisation des documents dans les repères ainsi constitués pour déterminer quels traits permettent des contrastes nets entre types de textes .

Articuler traits linguistiques / signalétiques pour examiner les corrélations entre les regroupements obtenus sur la base de traits linguistiques et les catégories documentaires existant par ailleurs [19].

[18] Par extraction de sous-matrices, par exemple.

[19] Comme les « genres » distingués par la documentation du *Monde* : interview, nécrologie, etc.

Combiner des traits Obtenir un « grain » plus gros.

Les tâches à entreprendre sont les suivantes :

Utilisation d'autres étiqueteurs Tree Tagger [20] ou Cordial 6 Universités ;

Amélioration des traits typologiques utilisés Les expériences présentes et passées (Illouz *et al.*, 1999) montrent les angles morts de la grille actuelle ;

Intégration logicielle Les phases de traitement ne sont pas encore articulées dans une architecture cohérente comme LT NSL ;

Structuration des traits Elle doit permettre une utilisation à « géométrie variable », allant de catégories très grossières (la partie du discours) à des étiquettes très fines ($Il\ faut_{Indicatif\ présent}$) voire à des regroupements transversaux (le genre ou le nombre, par exemple) ;

Mise en évidence de corrélations de traits Les techniques exploratoires vont être mises à contribution sur ce point.

Un nouveau projet, TyPWeb, dans le cadre d'une collaboration avec le CNET, vise à adapter pour le traitement de sites Web l'architecture mise en œuvre dans TyPTex et va nous conduire à passer du prototype actuel à une architecture générique de profilage. Ce projet vise à fournir un cadre méthodologique et pratique de profilage de sites Web et une typologie fine de ces sites. La démarche suivie vise à caractériser chaque site par des indicateurs de contenu et de structure. Il s'agit dans un premier temps de définir puis d'enrichir la description de sites par des indicateurs décrivant la forme et le contenu des sites visés : ces informations alimentent le cartouche descriptif des sites analysés, ce cartouche est conçu pour rester ouvert à toute nouvelle information pertinente capable de l'enrichir. TyPWeb doit conduire ensuite à proposer une typologie des contenus (en utilisant des index thématiques prédéfinis ou en constituant de nouvelles catégories de contenu en suivant la démarche inductive propre à l'architecture). L'analyse visée doit passer par un croisement de la structure formelle et des typologies de contenus produites. Il s'agit aussi de décrire l'articulation entre la description formelle et sémantique des sites avec les récits des pratiques des acteurs (concepteurs et visiteurs). Cette démarche vise en particulier à analyser la mise en place progressive de règles implicites d'échanges sur l'hypertoile.

Références

Adda, G., Mariani, J., Paroubek, P. & Lecomte, J. (1999). Métrique et premiers résultats de l'évaluation GRACE des étiqueteurs morphosyntaxiques pour le français. In P. Amsili, Ed., *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 15–24, Cargèse : ATALA.

S. Amstrong, Ed. (1994). *Using Large Corpora.* Cambridge, Massachusetts : The MIT Press.

Biber, D. (1988). *Variation accross speech and writing.* Cambridge : Cambridge University Press.

Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational Linguistics*, **19**(2), 243–258.

[20] http ://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html

Biber, D. (1995). *Dimensions of register variation : a cross-linguistic comparison.* Cambridge : Cambridge University Press.

Bronckart, J.-P., Bain, D., Schneuwly, B., Davaud, C. & Pasquier, A. (1985). *Le fonctionnement des discours : un modèle psychologique et une méthode d'analyse.* Lausanne : Delachaux & Niestlé.

Brunet, E. (1981). *Le vocabulaire français de 1789 à nos jours, d'après les données du Trésor de la langue française*, volume I of *Travaux de linguistique quantitative.* Genève/Paris : Slatkine/Champion. Préface de Paul Imbs.

Christ, O. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94 (3rd Conference on Computational Lexicography and Text Research)*, Budapest, Hungary. CMP-LG archive id 9408005.

Constant, P. (1991). *Analyse syntaxique par couches.* Doctorat de l'enst, École Nationale Supérieure des Télécommunications, Paris.

Dunlop, D. (1995). Practical considerations in the use of TEI headers in large corpora. *Computers and the Humanities*, (29), 85–98. Text Encoding Initiative. Background and Context, edited by Nancy Ide and Jean Véronis.

Gazdar, G., Pullum, G. K., Carpenter, R., Klein, E., Hukari, T. E. & Levine, R. D. (1990). Les structures de catégories. In P. Miller & T. Torris, Eds., *Formalismes syntaxiques pour le traitement automatique du langage naturel*, Langue, raisonnement, calcul, chapter 6, pp. 245–301. Paris : Hermès.

Habert, B., Nazarenko, A. & Salem, A. (1997). *Les linguistiques de corpus.* U Linguistique. Paris : Armand Colin/Masson.

Habert, B. & Salem, A. (1995). L'utilisation de catégorisations multiples pour l'analyse quantitative de données textuelles. *TAL*, **36**(1–2), 249–276. Traitements probabilistes et corpus, Benoît Habert (resp.).

Heiden, S. (1999). Encodage uniforme et normalisé de corpus. Application à l'étude d'un débat parlementaire. *Mots*, (60), 113–132. Presses de Sciences Po.

Heiden, S., Cuq, A., Ducout, D., Horlaville, P., Robert, J.-P., Prieur, V. & Dohm, B. (1998). *CorTeCs – 1.0β : Manuel de l'utilisateur.* Laboratoire de Lexicométrie et Textes Politiques – UMR 9952, CNRS – ENS Fontenay/Saint-Cloud.

N. Ide & J. Véronis, Eds. (1995). *The Text Encoding Initiative : Background and context.* Dordrecht : Kluwer Academic Publishers.

Illouz, G. (1999). Méta-étiqueteur adaptatif : vers une utilisation pragmatique des ressources linguistiques. In P. Amsili, Ed., *Actes de TALN'99 (Traitement Automatique des Langues Naturelles)*, pp. 15–24, Cargèse : ATALA.

Illouz, G., Habert, B., Fleury, S., Folch, H., Heiden, S. & Lafon, P. (1999). Maîtriser les déluges de données hétérogènes. In A. Condamines, C. Fabre & M.-P. Péry-Woodley, Eds., *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, pp. 37–46, Cargèse.

Ingenia (1995). *Manuel de développement Sylex-Base.* Ingenia – Langage naturel, Paris. 1.5.D.

Karlgren, J. (1999). Stylistic experiments in information retrieval. In T. Strzalkowski, Ed., *Natural Language Information Retrieval*, pp. 147–166. Pays-Bas : Kluwer.

Lafon, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. *MOTS*, (1), 128–165. Presses de la Fondation Nationale des Sciences Politiques.

McKelvie, D., Brew, C. & Thompson, H. (1997). Using SGML as a basis for data-intensive NLP. In *Proceedings 5th Conference on Applied NLP*, pp. 229–236 : ACL.

Naulleau, E. (1998). *Tranformation of* Le Monde *data to obtain PAROLE DTD conformance.* Technical report, INaLF – CNRS, Saint-Cloud.

Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computing*, (18), 401–409.

Sekine, S. (1998). The domain dependence of parsing. In *Fifth Conference on Applied Natural Language Processing*, pp. 96–102, Washington : Association for Computational Linguistics.

Sueur, J.-P. (1982). Pour une grammaire du discours : élaboration d'une méthode ; exemples d'application. *MOTS*, (5), 145–185.

Wilks, Y. & Gaizauskas, R. (1999). Lasie jumps the GATE. In T. Strzalkowski, Ed., *Natural language information retrieval*, Text, speech and language technology, chapter 8, pp. 197–214. Dordrecht : Kluwer.